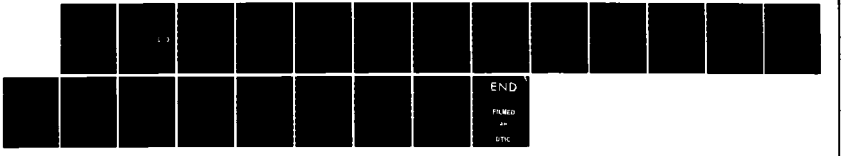
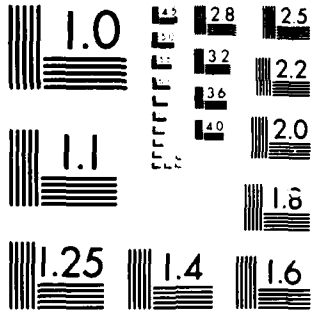
AD-A163 346 ASYMPTOTIC NORMALITY OF U-STATISTICS BASED ON TRIMMED SAMPLES(U) JOHNS HOPKINS UNIV BALTIMORE MD DEPT OF MATHEMATICAL SCIENCES P JANSSEN ET AL. NOV 85 TR-450 1/1

UNCLASSIFIED ONR-85-5 N00014-79-C-0801 F/G 12/1 NL





DEPARTMENT OF MATHEMATICAL SCIENCES
The Johns Hopkins University
Baltimore, Maryland 21218

ASYMPTOTIC NORMALITY OF U-STATISTICS BASED ON TRIMMED SAMPLES

by

Paul Janssen[1], Robert Serfling[2], and Noël Veraverbeke[1]

Technical Report No. 450

ONR Technical Report No. 85-5

November, 1985

[1]Limburgs Universitair Centrum, Diepenbeek, Belgium

[2]Johns Hopkins University, Baltimore, U.S.A.

Research supported by the U.S. Department of Navy under Office of Naval Research Contract No. N00014-79-C-0801. Reproduction in whole or in part is permitted for any purpose of the United States Government.

ABSTRACT

ASYMPTOTIC NORMALITY OF U-STATISTICS BASED
ON TRIMMED SAMPLES

Let $X_{n1} \leq \ldots \leq X_{nn}$ be an ordered sample of size n. We establish asymptotic normality of U-statistics based on the trimmed sample $X_{n,[\alpha n]+1} \leq \ldots \leq X_{n,n-[\beta n]}$, where $0 < \alpha,\beta < 1/2$. This theorem and its multi-sample generalization are illustrated by various statistics of importance for robust estimation of location, dispersion, etc.

This unifies the flexibility of the class of U-statistics and the classical principle of rejection of outliers. In addition, as a tool in our treatment, but also having broader interest, a uniform version of the central limit theorem for U-statistics is provided.

1. Introduction. For robust estimation of location, the ordinary sample mean is too sensitive to outliers. A classical and successful alternative is the trimmed mean, for which asymptotic normality was established by Bickel (1965). As discussed by Bickel and Lehmann (1975), for example, the trimmed mean remains relatively efficient with respect to the untrimmed mean even in the absence of outliers.

Viewing the trimmed mean as simply the ordinary mean defined on a trimmed sample, we are motivated to consider other common statistics as well in this regard. In this paper we study U-statistics in such fashion. The class of U-statistics, introduced by Hoeffding (1948), contains a wealth of statistics of interest in their own rights and also contains statistics which serve as approximations to statistics of more complicated type. A very significant broadening of the scope of robust statistical inference is achieved, therefore, by consideration of the class of U-statistics on trimmed samples.

Specifically, the statistics we treat are defined as follows. Let $X_1,\ldots,X_n$ be an i.i.d. sample from a df F, let $X_{n1} \leq \ldots \leq X_{nn}$ denote the ordered X_i's, let $0 < \alpha,\beta < 1/2$, and put $n_{\alpha\beta} = n - [\alpha n] - [\beta n]$. Let $h(x_1,\ldots,x_m)$ be a "kernel" assumed (without loss of generality) to be symmetric in its arguments. For each such kernel we consider the associated U-statistic defined on the (α,β)-trimmed sample, i.e.,

$$U_{n\alpha\beta} = \binom{n_{\alpha\beta}}{m}^{-1} \sum_{C_{n\alpha\beta}} h(X_{ni_1},\ldots,X_{ni_m}), \tag{1.1}$$

where $C_{n\alpha\beta}$ denotes the set of m-tuples $\{(i_1,\ldots,i_m): [\alpha n]+1 \leq i_1 < \ldots < i_m \leq n-[\beta n]\}$. (The cases $\alpha > 0$, $\beta = 0$ and $\alpha = 0$, $\beta > 0$ could also be considered but will be omitted for simplicity. The case $\alpha = \beta = 0$ corresponds to ordinary U-statistics based on full samples, for which there is already an extensive literature (see, e.g., Serfling (1980), Chapter 5)).

AMS Subject Classifications : Primary 62E20, Secondary 62G35

Key words and phrases : U-statistics, trimmed samples, robust inference, nonparametric, uniform central limit theory.

For $m = 1$ and the kernel $h(x) = x$, (1.1) gives the trimmed mean

$$(1.2) \qquad \bar{X}_{n\alpha\beta} = n_{\alpha\beta}^{-1} \sum_{i=[\alpha n]+1}^{n-[\beta n]} X_{ni} ,$$

treated by Bickel (1965). For $m = 2$ and the kernel $h(x_1,x_2) = I\{x_1+x_2>0\}$ we obtain a version of the Wilcoxon one-sample statistic treated by Saleh (1976). Evidently, these and only one or two other cases of (1.1) have been previously studied in the literature, despite the abundance of natural possibilities. For example, with $m = 2$ and $h(x_1,x_2) = (x_1-x_2)^2/2$, we obtain the statistic

$$(1.3) \qquad S_{n\alpha\beta}^2 = (n_{\alpha\beta}-1)^{-1} \sum_{i=[\alpha n]+1}^{n-[\beta n]} (X_{ni} - \bar{X}_{n\alpha\beta})^2 ,$$

a quite natural robust analogue of the classical sample variance. Surprisingly, this has not been examined previously, although a Winsorized version has been treated by Jaeckel (1971) (see also Bickel and Doksum (1977), p. 375).

Other methods of using trimming to produce robustness have appeared in the literature. For example, see the "doubly trimmed standard deviation" of Bickel and Lehmann (1976), and a "trimmed standard deviation" introduced by Bickel and Lehmann (1979) and treated theoretically by Janssen, Serfling and Veraverbeke (1984). Also, consider the "trimmed U-statistics", produced by trimming on the basis of ordered values of $h(X_{i_1},\ldots,X_{i_m})$, which are a special case of the generalized L-statistics treated by Serfling (1984). The present development, in which the trimming is applied directly to the sample values X_i, is perhaps the most natural and reasonable way to implement a principle of rejection of outliers.

In Section 2 we establish asymptotic normality for statistics of form (1.1) (under suitable regularity conditions), thereby extending and unifying Bickel's result for the trimmed mean and Hoeffding's result for the case of untrimmed samples. We also cover the multi-sample case. The main results are given by Theorems 2.1 and 2.2.

Our method of proof utilizes recent work of Randles (1982) on U-statistics based on kernels having unknown parameters. However, an alternate approach consists of extension of Bickel's method for the trimmed mean. This is discussed in Remark B of Section 2. The approach entails proving a result which is of general interest, a uniform central limit theorem for U-statistics, extending Parzen (1954). We present this result in the Appendix.

The remainder of the paper (Section 3) treats important examples : moment-type statistics, Wilcoxon-type statistics and Gini-like measures of location and spread. A number of interesting new statistics are examined.

2. General theorems. Our asymptotic normality result for the statistic $U_{n\alpha\beta}$ defined by (1.1) will involve mean and variance parameters $\mu_{\alpha\beta}$ and $\sigma^2_{\alpha\beta}$ defined as follows. Corresponding to the given kernel h, we define for $x,u,v \in \mathbb{R}$

$$g(x;u,v) = (1-\beta-\alpha)^{-m} I\{u \le x \le v\} \int_u^v \cdots \int_u^v h(x,x_1,\ldots,x_{m-1}) \prod_{i=1}^{m-1} dF(x_i),$$

$$\mu(u,v) = Eg(X;u,v) = (1-\beta-\alpha)^{-m} \int_u^v \cdots \int_u^v h(x_1,\ldots,x_m) \prod_{i=1}^{m} dF(x_i),$$

$$\Delta(u,v) = \mathrm{Var}\, g(X;u,v),$$

$$A(u,v) = -g(u;u,v),$$

and

$$B(u,v) = g(v;u,v).$$

Then we define

$$\mu_{\alpha\beta} = \mu(F^{-1}(\alpha), F^{-1}(1-\beta))$$

and

$$\sigma^2_{\alpha\beta} = m^2\{\Delta_{\alpha\beta} + 2\alpha\mu_{\alpha\beta} A_{\alpha\beta} - 2\beta\mu_{\alpha\beta} B_{\alpha\beta} + \alpha(1-\alpha) A^2_{\alpha\beta}$$

$$+ \beta(1-\beta) B^2_{\alpha\beta} + 2\alpha\beta A_{\alpha\beta} B_{\alpha\beta}\},$$

where $\Delta_{\alpha\beta} = \Delta(F^{-1}(\alpha), F^{-1}(1-\beta))$, etc.

ASSUMPTIONS.

(A) F has a density f which is continuous and positive at $F^{-1}(\alpha)$ and $F^{-1}(1-\beta)$ and is bounded in some δ-neighborhoods of $F^{-1}(\alpha)$ and $F^{-1}(1-\beta)$;

(B) For some $a < F^{-1}(\alpha)$ and $b > F^{-1}(1-\beta)$:

$$\sup_{a \le x_1,\ldots,x_m \le b} |h(x_1,\ldots,x_m)| = M_o < \infty;$$

(C) The function $g(x) = \int_{F^{-1}(\alpha)}^{F^{-1}(1-\beta)} \cdots \int_{F^{-1}(\alpha)}^{F^{-1}(1-\beta)} h(x,x_1,\ldots,x_{m-1}) \prod_{i=1}^{m-1} dF(x_i)$ is continuous at $F^{-1}(\alpha)$ and $F^{-1}(1-\beta)$.

THEOREM 2.1. Let $U_{n\alpha\beta}$ be given by (1.1). Assume conditions (A), (B) and (C) and that $\sigma^2_{\alpha\beta} > 0$. Then

$$n^{1/2}\,(U_{n\alpha\beta} - \mu_{\alpha\beta}) \xrightarrow{d} N(0,\sigma^2_{\alpha\beta}).$$

The proof will utilize a series of lemmas involving certain U-statistics closely related to $U_{n\alpha\beta}$. For the given kernel h, and for each $u < v \in \mathbb{R}$, we define an associated kernel

$$h(x_1,\ldots,x_m;u,v) = (1-\beta-\alpha)^{-m}\, I\{u \le x_1,\ldots,x_m \le v\}\, h(x_1,\ldots,x_m)$$

and we denote by $U_n(u,v)$ the ordinary (i.e., defined on the full sample) U-statistic based on this kernel.

Then we have

$$U_{n\alpha\beta} = \binom{n}{m\alpha\beta}^{-1} \binom{n}{m} (1-\beta-\alpha)^m\, U_n(X_{n,[\alpha n]+1}, X_{n,n-[\beta n]}) \tag{2.1}$$

and we readily obtain

LEMMA 2.1. Under condition (B),

$$U_{n\alpha\beta} = U_n(X_{n,[\alpha n]+1}\ ,\ X_{n,n-[\beta n]}) + O_p(n^{-1}). \tag{2.2}$$

Next we show that the leading term in (2.2) may be approximated by

$$T_{n\alpha\beta} = U_n(F^{-1}(\alpha),\ F^{-1}(1-\beta)) - \mu_{\alpha\beta} + \mu(X_{n,[\alpha n]+1}\ ,\ X_{n,n-[\beta n]}).$$

LEMMA 2.2. Under conditions (A) and (B),

$$T_{n\alpha\beta} = U_n(X_{n,[\alpha n]+1}\ ,\ X_{n,n-[\beta n]}) + o_p(n^{-1/2}). \tag{2.3}$$

PROOF. We make direct application of Theorem 2.8 of Randles (1982), by which (2.3) holds if Randles' Conditions 2.2 and 2.3 are fulfilled. First we note that, by classical central limit theory for order statistics (or by (2.7) and (2.8) below),

$$(X_{n,[\alpha n]+1}\ ,\ X_{n,n-[\beta n]}) - (F^{-1}(\alpha),\ F^{-1}(1-\beta)) = O_p(n^{-1/2}). \tag{2.4}$$

This is Randles' Condition 2.2 specialized to our setting. Next we note that by condition (B) there exists $M_1 < \infty$ such that for all $x_1,\ldots,x_m$ and all (u,v) in some neighborhood of $(F^{-1}(\alpha),\ F^{-1}(1-\beta))$,

$$\left| h(x_1,\ldots,x_m;u,v) - h(x_1,\ldots,x_m;F^{-1}(\alpha),\ F^{-1}(1-\beta)) \right| \leq M_1. \tag{2.5}$$

Now let K be a neighborhood of $(F^{-1}(\alpha),\ F^{-1}(1-\beta))$ which is contained in the rectangular neighborhood of $(F^{-1}(\alpha),\ F^{-1}(1-\beta))$ in which (A) and (B) hold. For $(u,v) \in K$ and for a sphere D centered at (u,v) with radius d, such that $D \subset K$, we have

$$\sup_{(u',v') \in D} \left| h(x_1,\ldots,x_m;u',v') - h(x_1,\ldots,x_m;u,v) \right|$$

$$\leq (1-\beta-\alpha)^{-m} \left| h(x_1,\ldots,x_m) \right| \Big[\sum_{i=1}^{m} I\{u-d \leq x_i \leq u+d\} \prod_{j \neq i} I\{u-d \leq x_j \leq v+d\}$$

$$+ \sum_{i=1}^{m} I\{v-d \leq x_i \leq v+d\} \prod_{j \neq i} I\{u-d \leq x_j \leq v+d\} \Big].$$

Then, using (A) and (B), we obtain

$$E\Big\{ \sup_{(u',v') \in D} \left| h(X_1,\ldots,X_m;u',v') - h(X_1,\ldots,X_m;u,v) \right| \Big\}$$

$$\leq m(1-\beta-\alpha)^{-m} \Big[\Big(\int_{u-d}^{u+d} \int_{u-d}^{v+d} \cdots \int_{u-d}^{v+d} + \int_{v-d}^{v+d} \int_{u-d}^{v+d} \cdots \int_{u-d}^{v+d} \Big) \left| h \right| \prod_{i=1}^{m} dF(x_i) \Big]$$

$$\leq 2m\, M_0 (1-\beta-\alpha)^{-m} [F(v+d)-F(u-d)]^{m-1} \{[F(u+d)-F(u-d)]+[F(v+d)-F(v-d)]\}$$

$$\leq M_2 d, \tag{2.6}$$

for suitable choice of constant M_2 not depending on choice of D. By Randles' Lemma 2.6, his Condition 2.3 follows from our (2.5) and (2.6). Thus our lemma follows. □

LEMMA 2.3. Under condition (B),

$$U_n(F^{-1}(\alpha), F^{-1}(1-\beta)) - \mu_{\alpha\beta} = \frac{m}{n} \sum_{i=1}^{n} [g(X_i;F^{-1}(\alpha),F^{-1}(1-\beta)) - \mu_{\alpha\beta}]$$

$$+ o_p(n^{-1/2}).$$

PROOF. This is immediate from the projection theory of ordinary U-statistics (see, e.g., Serfling (1980), Chapter 5). □

LEMMA 2.4. Under condition (A) and (C),

$$\mu(X_{n,[\alpha n]+1}\ ,\ X_{n,n-[\beta n]}) - \mu_{\alpha\beta}$$

$$= mA_{\alpha\beta}\,[\alpha - F_n(F^{-1}(\alpha))] + mB_{\alpha\beta}\,[1-\beta-F_n(F^{-1}(1-\beta))] + o_p(n^{-1/2}),$$

where F_n denotes the usual sample df of $X_1,\ldots,X_n$.

PROOF. By condition (A) and a result of Ghosh (1971) on Bahadur representation of order statistics, we have

$$X_{n,[\alpha n]+1} - F^{-1}(\alpha) = \frac{\alpha - F_n(F^{-1}(\alpha))}{f(F^{-1}(\alpha))} + o_p(n^{-1/2}) \tag{2.7}$$

and

$$X_{n,n-[\beta n]} - F^{-1}(1-\beta) = \frac{1-\beta-F_n(F^{-1}(1-\beta))}{f(F^{-1}(1-\beta))} + o_p(n^{-1/2}). \tag{2.8}$$

By conditions (A) and (B) again along with condition (C),

$$\left.\frac{\partial\mu}{\partial u}\right|_{(u,v)=(F^{-1}(\alpha),F^{-1}(1-\beta))} = m\,f(F^{-1}(\alpha))A_{\alpha\beta}$$

and

$$\left.\frac{\partial\mu}{\partial v}\right|_{(u,v)=(F^{-1}(\alpha),F^{-1}(1-\beta))} = m\,f(F^{-1}(1-\beta))B_{\alpha\beta}.$$

Thus, by the multivariate version of Young's form of Taylor's theorem (e.g., an immediate extension of Theorem C on page 45 of Serfling (1980)),

$$\begin{aligned}\mu(u,v) - \mu_{\alpha\beta} &= m\,f(F^{-1}(\alpha))\,A_{\alpha\beta}\,(u-F^{-1}(\alpha)) \\ &\quad + m\,f(F^{-1}(1-\beta))\,B_{\alpha\beta}\,(v-F^{-1}(1-\beta)) \\ &\quad + o(\|\,(u,v) - (F^{-1}(\alpha),\,F^{-1}(1-\beta))\,\|\,).\end{aligned} \tag{2.9}$$

Applying (2.4), (2.7) and (2.8) in (2.9), we obtain the desired result. □

PROOF OF THEOREM 2.1. Define

$$\psi(x;u,v) = m\{[g(x;u,v) - \mu(u,v)] + A(u,v)[F(u) - I\{x\le u\}] + B(u,v)[F(v) - I\{x\le v\}]\}.$$

Then, combining Lemmas 2.1 - 2.4, we may write

$$U_{n\alpha\beta} - \mu_{\alpha\beta} = n^{-1}\sum_{i=1}^{n}\psi(X_i;F^{-1}(\alpha),F^{-1}(1-\beta)) + o_p(n^{-1/2}).$$

Finally, note that $\psi(X_1;F^{-1}(\alpha), F^{-1}(1-\beta))$ has mean 0 and variance $\sigma^2_{\alpha\beta}$. □

REMARK A. By Lemmas 2.1 and 2.2, we may write

$$U_{n\alpha\beta} - \mu_{\alpha\beta} = U_n^* + O_n^* + o_p(n^{-1/2}), \tag{2.10}$$

where U_n^* is an ordinary U-statistic with mean 0 and asymptotic variance parameter $m^2\Delta_{\alpha\beta}$, and O_n^* is a function of (two) order statistics. It can happen that $\sigma^2_{\alpha\beta} > 0$ but one of these two components is negligible, namely U_n^* if $\Delta_{\alpha\beta} = 0$ and O_n^* if $A_{\alpha\beta} = B_{\alpha\beta} = 0$. The case that U_n^* is <u>non-negligible</u> (i.e., $\Delta_{\alpha\beta} > 0$) can occur even when the ordinary U-statistic based on the <u>original</u> kernel h is degenerate. That is, a U-statistic which has nonnormal limit distribution when defined on the full sample can have a normal limit distribution when defined on a trimmed sample. □

It is straightforward to extend Theorem 2.1 to the case of multi-sample U-statistics. For simplicity, we consider the 2-sample situation. Let $X_1,\ldots,X_{n_1}$ be an i.i.d. sample from df F_1 and $Y_1,\ldots,Y_{n_2}$ i.i.d. from F_2. Let $h(x_1,\ldots,x_{m_1}; y_1,\ldots,y_{m_2})$ be symmetric within blocks and consider

$$(2.11) \qquad U_{\underset{\sim}{n}\underset{\sim}{\alpha}\underset{\sim}{\beta}} = \prod_{i=1}^{2} \binom{n_i - [\alpha_i n_i] - [\beta_i n_i] - 1}{m_i} \cdot$$

$$\sum_{C_{\underset{\sim}{n}\underset{\sim}{\alpha}\underset{\sim}{\beta}}} h\,(X_{n_1 i_1},\ldots,X_{n_1 i_{m_1}}\,;\,Y_{n_2 j_1},\ldots,Y_{n_2 j_{m_2}})$$

where $C_{\underset{\sim}{n}\underset{\sim}{\alpha}\underset{\sim}{\beta}} = \{(i_1,\ldots,i_{m_1};j_1,\ldots,j_{m_2}) : [\alpha_1 n_1]+1 \le i_1 < \ldots < i_{m_1} \le n_1 - [\beta_1 n_1]$

and $[\alpha_2 n_2]+1 \le j_1 < \ldots < j_{m_2} \le n_2 - [\beta_2 n_2]\}$.

For $u<v \in \mathbb{R}$ and $u'<v' \in \mathbb{R}$, define associated kernels

$$h(x_1,\ldots,x_{m_1};y_1,\ldots,y_{m_2};\,u,v;u',v')$$

$$= \prod_{i=1,2} (1-\beta_i-\alpha_i)^{-m_i} h(x_1,\ldots,x_{m_1};y_1,\ldots,y_{m_2}).$$

$$I\{u \le x_1,\ldots,x_{m_1} \le v, u' \le y_1,\ldots,y_{m_2} \le v'\}.$$

Denote by $g_1(x;u,v,u',v')$ the conditional expectation of this kernel given $X_1=x$, and by $g_2(y;u,v,u',v')$ the conditional expectation given $Y_1=y$. Put

$$\mu(u,v,u',v') = Eg_1(X_1;u,v,u',v') = Eg_2(Y_1;u,v,u'v'),$$

$$\Delta_1(u,v,u',v') = \mathrm{Var}\, g_1(X_1;u,v,u',v'),$$

and

$$\Delta_2(u,v,u',v') = \mathrm{Var}\, g_2(Y_1;u,v,u',v').$$

Let μ, Δ_1, Δ_2 denote the evaluations of these quantities at $(F_1^{-1}(\alpha_1), F_1^{-1}(1-\beta_1), F_2^{-1}(\alpha_2), F_2^{-1}(1-\beta_2))$.
Further let $\tilde{A}_1$, $\tilde{B}_1$, $\tilde{A}_2$, $\tilde{B}_2$ denote the partial derivatives of $\mu(u,v,u',v')$ w.r.t. u,v,u' and v' respectively, evaluated at $(F_1^{-1}(\alpha_1), F_1^{-1}(1-\beta_1), F_2^{-1}(\alpha_2), F_2^{-1}(1-\beta_2))$.

Set $A_i = \tilde{A}_i/m_i f_i(F_i^{-1}(\alpha_i))$ and $B_i = \tilde{B}_i/m_i f_i(F_i^{-1}(1-\beta_i))$ for i=1,2, assume that

$$\frac{n_i}{n_1+n_2} \to \lambda_i \quad (i=1,2) \qquad \text{as } \min(n_1,n_2) \to \infty,$$

and define

$$\sigma^2 = \sum_{i=1,2} \frac{m_i^2}{\lambda_i} [\Delta_i + 2\alpha_i \mu A_i - 2\beta_i \mu B_i + \alpha_i(1-\alpha_i)A_i^2$$

$$+ \beta_i(1-\beta_i)B_i^2 + 2\alpha_i\beta_i A_i B_i].$$

Then we have :

THEOREM 2.2. Assume that F_i (i=1,2) and h satisfy (analogues of) conditions (A), (B), (C) and assume $\sigma^2 > 0$. Then, as $\min(n_1,n_2) \to \infty$ such that $n_i/(n_1+n_2) \to \lambda_i (i=1,2)$,

$$(n_1+n_2)^{1/2} (U_{n\underline{\alpha}\underline{\beta}} - \mu) \xrightarrow{d} N(0,\sigma^2).$$

The proof is similar to that of Theorem 2.1. Also, from the above formula for the asymptotic variance, it is easy to recognize what the asymptotic variance parameter should be in the c-sample case for c > 2.

REMARK B. Bickel (1965), in treating the trimmed mean, used a different method of proof, which was also adapted and followed by Saleh (1976) with some gaps in the development. Such an approach also can be followed for our Theorem 2.1, as follows. Conditional on $X_{n,[\alpha n]}$ and $X_{n,n-[\beta n]+1}$, the statistic $U_{n\alpha\beta}$ is distributed as an ordinary U-statistic based on an i.i.d. sample of size $n_{\alpha\beta}$ from a certain df depending on the given two order statistics. By a uniform CLT for U-statistics (see Appendix), we can apply Theorem 2 of Sethuraman (1961) and complete the proof in the manner of Bickel (1965). □

3. EXAMPLES.

3.1. Central moments

Here we consider robust (i.e., trimmed sample) versions of the classical measures of location, dispersion, skewness, kurtosis, etc. Since all central moments may be represented as U-statistics (Hoeffding (1948), p. 295), Theorem 2.1 yields the appropriate results. In particular, let us symmetrically trim the sample ($\alpha=\beta$) and consider the trimmed mean (1.2) and the trimmed variance (1.3). Let us also confine attention to df's which are symmetric about 0. Then Theorem 1.2 yields

$$n^{1/2}\,\bar{X}_{n\alpha\alpha} \xrightarrow{d} N(0,\sigma_1^2(\alpha,1-\alpha)), \tag{3.1}$$

with

$$\sigma_1^2(\alpha,1-\alpha) = (1-2\alpha)^{-2}\left[\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x^2 f(x)dx + 2\alpha(F^{-1}(\alpha))^2\right], \tag{3.2}$$

which corresponds to Bickel (1965), and

$$n^{1/2}\,S_{n\alpha\alpha}^2 \xrightarrow{d} N(\mu(\alpha,1-\alpha),\ \sigma_2^2(\alpha,1-\alpha)), \tag{3.3}$$

with

$$\mu(\alpha,1-\alpha) = (1-2\alpha)^{-1}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x^2 f(x)dx \tag{3.4}$$

and

$$\begin{aligned}\sigma_2^2(\alpha,1-\alpha) = (1-2\alpha)^{-2}\Big[&\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x^4 f(x)dx - \Big(\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x^2 f(x)dx\Big)^2 \\ &- 4\alpha(F^{-1}(\alpha))^2\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x^2 f(x)dx + 2\alpha(1-\alpha)(F^{-1}(\alpha))^4\Big].\end{aligned} \tag{3.5}$$

This explicit result for the asymptotic variance makes it possible to compare $S_{n\alpha\alpha}^2$ with competitors such as the usual sample variance and the mean absolute deviation, using an appropriate asymptotic relative efficiency criterion based on asymptotic variance parameters.

3.2. Wilcoxon-like statistics

In the one-sample case, assume F has density f symmetric about Δ and consider testing $\Delta = 0$ versus $\Delta > 0$. This can be formulated as a problem of testing $P\{X_1 + X_2 > 0\} = 1/2$ versus $P\{X_1 + X_2 > 0\} > 1/2$, with corresponding test statistic

$$U_n = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} I\{X_i + X_j > 0\},$$

which is asymptotically equivalent to the one-sample Wilcoxon statistic. Trimming leads to consideration of

$$U_{n\alpha\alpha} = \binom{n-2[\alpha n]}{2}^{-1} \sum_{[\alpha n]+1 \le i < j \le n-[\alpha n]} I\{X_{ni} + X_{nj} > 0\}.$$

From Theorem 2.1 it follows by routine calculations that, under the null hypothesis $\Delta = 0$,

$$n^{1/2}\,(U_{n\alpha\alpha} - 1/2) \xrightarrow{d} N\left(0, \frac{1+4\alpha}{3(1-2\alpha)^2}\right), \tag{3.6}$$

a result previously found by Saleh (1976).

In the two-sample case, let $P = \{(F,G) : F(x) \le G(x), \text{ all } x\}$ and consider testing $F = G$ versus $F \ne G$, or in turn consider testing $P\{X_1 < Y_1\} = 1/2$ versus $P\{X_1 < Y_1\} < 1/2$. The usual Wilcoxon-Mann-Whitney statistic has the following formulation in the trimmed-sample case :

$$U_{n_1 n_2 \alpha} = \frac{1}{(n_1 - 2[\alpha n_1])(n_2 - 2[\alpha n_2])} \sum_{i=[\alpha n_1]+1}^{n_1-[\alpha n_1]} \sum_{j=[\alpha n_2]+1}^{n_2-[\alpha n_2]} I\{X_{n_1 i} < Y_{n_2 j}\},$$

which can be studied by our Theorem 2.2, which yields

$$n^{1/2}\,(U_{n_1 n_2 \alpha} - \tfrac{1}{2}) \xrightarrow{d} N\left(0, \frac{1}{\lambda(1-\lambda)} \cdot \frac{1+4\alpha}{12(1-2\alpha)^2}\right) \tag{3.7}$$

under the null hypothesis, where $n = n_1 + n_2$ and $\lambda = \lim n_1/n$.
This result can also be found in Hettmansperger (1968).

3.3. Gini-like measures of location and dispersion

For estimating location, Heilmann (1980) proposed to modify the sample mean, which can be expressed as

$$\bar{X}_n = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \text{median } (X_i, X_j)$$

to

$$L_n = \binom{n}{3}^{-1} \sum_{1 \le i < j < k \le n} \text{median } (X_i, X_j, X_k).$$

In similar vein, we consider as competitors of the trimmed sample mean the statistics $U_{n\alpha\beta}$ of the form (1.1) corresponding to the kernels given by

$$h(x_1, \ldots, x_m) = \text{median } (x_1, \ldots, x_m) \tag{3.8}$$

for $m = 3,4,\ldots$ Asymptotic normality is provided by Theorem 2.1. Let us note that naive computation of the statistic $U_{n\alpha\beta}$ corresponding to (3.8) would require $O(n^m)$ steps. It is of interest, therefore, that for the kernel (3.8) $U_{n\alpha\beta}$ may also be represented as an L-statistic,

$$U_{n\alpha\beta} = \sum_{i=[\alpha n]+1}^{n-[\beta n]} c_{ni} X_{ni}, \tag{3.9}$$

where, with $\binom{a}{b} = 0$ if $a < b$,

$$c_{ni} = \frac{1}{2} \binom{n}{m}^{-1}_{\alpha\beta} \left[\binom{i-[\alpha n]-1}{[\frac{m-1}{2}]} \binom{n-[\beta n]-i}{[\frac{m}{2}]} + \binom{i-[\alpha n]-1}{[\frac{m}{2}]} \binom{n-[\beta n]-i}{[\frac{m-1}{2}]} \right]. \tag{3.10}$$

Thus, given the c_{ni}'s (which itself is a computational problem, of course), one can compute the statistic via the formula (3.9) in $O(n \log n)$ steps.

For estimating dispersion, Heilmann (1980) proposed to modify Gini's mean difference, which is the U-statistic based on kernel $h(x_1,x_2) = |x_1 - x_2|$, but which also can be expressed as

$$(3.11) \qquad G_n = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \text{range}\,(X_i, X_j)$$

to

$$(3.12) \qquad G_n^* = \binom{n}{3}^{-1} \sum_{1 \le i < j < k \le n} \text{range}\,(X_i, X_j, X_k).$$

In similar vein, we consider as competitors to G_n and G_n^* the class of statistics $U_{n\alpha\beta}$ of form (1.1) corresponding to the kernels given by

$$(3.13) \qquad h(x_1,\ldots,x_m) = \text{range}\,(x_1,\ldots,x_m)$$

for $m = 3,4,\ldots$. Asymptotic normality is provided by Theorem 2.1.

It is well-known that the statistic G_n may be written as an L-statistic (see, e.g. Serfling (1980)) and this is shown by Heilmann (1980) to be true also for G_n^*. It is easily seen that in fact, this is true in general for the statistics $U_{n\alpha\beta}$ based on the kernel (3.13), including the case $\alpha = \beta = 0$ (untrimmed sample). The relevant c_{ni}'s for a representation of form (3.9) are found to be

$$(3.14) \qquad c_{ni} = \binom{n}{m}_{\alpha\beta}^{-1} \left[\binom{i-[\alpha n]-1}{m-1} - \binom{n-[\beta n]-i}{m-1}\right].$$

An alternative way of generalizing the classical Gini's mean difference is by considering the class of U-statistcs based on the kernels of form

$$(3.15) \qquad h(x_1,x_2) = |x_1 - x_2|^p,$$

for $p > 0$. This gives the "p-th power measures" considered by Bickel and Lehmann (1979). However, results of Boos (1979) indicate that the case $p = 1$ is highly competitive to the cases $p \neq 1$. It is of interest to examine the class of "p-th power measures" on <u>trimmed</u> samples, which is now possible via Theorem 2.1.

Appendix : a uniform CLT for U-statistics

Here we provide a useful extension to U-statistics of the uniform CLT for sample means given by Parzen (1954). Let $X_1,\ldots,X_n$ be independent $\mathbb{R}$-valued random variables having common df $F(x;\theta)$, where $\theta \in \Theta \subset \mathbb{R}^p$, $p \geq 1$, and let U_n denote the usual U-statistic based on a given symmetric kernel $h(x_1,\ldots,x_m)$. Define

$$\mu(\theta) = E_\theta h,$$

$$g(x,\theta) = E_\theta\{h(X_1,\ldots,X_m) \mid X_1 = x\},$$

and

$$\sigma^2(\theta) = \mathrm{Var}_\theta g(X,\theta).$$

THEOREM. Assume

(1) $E_\theta h^2 \leq K < \infty$, all $\theta \in \Theta$,

and

(2) $\sigma^2(\theta) > L > 0$, all $\theta \in \Theta$,

for fixed constants K and L, and

(3) $\lim_{M\to\infty} E_\theta\,[[g(X,\theta) - \mu(\theta)]^2\, I\{|X| > M\}] = 0$, uniformly in $\theta \in \Theta$.

Then, uniformly in $\theta \in \Theta$,

$$\sup_x \left| P\{n^{1/2}\,(U_n-\mu(\theta))/m\sigma(\theta) \leq x\} - \Phi(x) \right| = o(1).$$

PROOF. Using the projection lemma as in Hoeffding (1948), we decompose $n^{1/2}\,(U_n-\mu(\theta))/m\sigma(\theta)$ into a leading term which is a properly normalized sum of i.i.d. r.v.'s and a remainder term. By a uniform version of Slutzky's Theorem (Parzen (1954), Theorem 18D), it suffices to show that, uniformly in $\theta \in \Theta$, the leading term has a normal limit and the remainder term tends to 0 in probability.

The former is obtained from the uniform normal convergence theorem for sample means (Parzen (1954), p. 38), while the latter follows by an application of Chebyshev's inequality. □

REFERENCES

Bickel, P.J. (1965), "On some robust estimates of location", Ann. Math. Statist., 36, 847-858.

Bickel, P.J. and Doksum, K.A. (1977), Mathematical Statistics, Holden-Day, San Francisco.

Bickel, P.J. and Lehmann, E.L. (1975), "Descriptive statistics for nonparametric models. II. Location", Ann. Statist., 3, 1045-1069.

Bickel, P.J. and Lehmann, E.L. (1976), "Descriptive statistics for nonparametric models. III. Dispersion", Ann. Statist., 4, 1139-1158.

Bickel, P.J. and Lehmann, E.L. (1979), "Descriptive statistics for nonparametric models. IV. Spread", in Contributions to Statistics. Hájek Memorial Volume (ed. by J. Jurečková), pp. 33-40, Academia, Prague.

Boos, D.D. (1979), "Gini's mean difference as a nonparametric measure of scale", Institute of Statistics Mimeo Series # 1166, North Carolina State University, Raleigh, North Carolina.

Ghosh, J.K. (1971), "A new proof of the Bahadur representation of quantiles and an application", Ann. Math. Statist., 42, 1957-1961.

Heilmann, W.R. (1980), "Basic distribution theory for nonparametric Gini-like measures of location and dispersion", Biometrical J., 22, 51-60.

Hettmansperger, T.P. (1968), "On the trimmed Mann-Whitney statistic", Ann. Math. Statist., 39, 1610-1614.

Hoeffding, W. (1948), "A class of statistics with asymptotically normal distribution", Ann. Math. Statist., 19, 293-325.

Jaeckel, L.A. (1971), "Robust estimates of location : symmetry and asymetric contamination", Ann. Math. Statist., 42, 1020-1034.

Janssen, P., Serfling, R. and Veraverbeke, N. (1984), "Asymptotic normality for a general class of statistical functions and applications to measures of spread", Ann. Statist., 12, 1369-1379.

Parzen, E. (1954), "On uniform convergence of families of sequences of random variables", Univ. of California Publ. Statist., 2, 23-54.

Randles, R.H. (1982), "On the asymptotic normality of statistics with estimated parameters", Ann. Statist., 10, 462-474.

Saleh, A.K. (1976). "Hodges-Lehmann estimate of the location parameter in censored samples", Ann. Inst. Statist. Math., 28, 235-247.

Serfling, R.J. (1980), Approximation Theorems of Mathematical Statistics, Wiley, New York.

Serfling, R.J. (1984), "Generalized L-, M- and R-statistics", Ann. Statist., 12, 76-86.

Sethuraman, J. (1961), "Some limit theorems for joint distributions", Sankhyā, Ser. A, 10, 379-386.

P. Janssen
N. Veraverbeke
Limburgs Universitair Centrum
Universitaire Campus
B-3610 Diepenbeek, Belgium

R.J. Serfling
Department of Mathematical Sciences
The Johns Hopkins University
Baltimore, Maryland 21218, U.S.A.

SECURITY CLASSIFICATION OF THIS PAGE

AD-A163 346

REPORT DOCUMENTATION PAGE

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT CATALOG NUMBER
ONR No. 85-5		

4. TITLE	5. TYPE OF REPORT & PERIOD COVERED
Asymptotic normality of U-statistics based on trimmed samples.	Technical Report
	6. PERFORMING ORGANIZATION REPORT NO. Technical Report No. 450

7. AUTHOR(s)	8. CONTRACT OR GRANT NUMBER(s)
Paul Janssen, Robert Serfling, Noel Veraverbeke	ONR No. N00014-79-C-0801

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Department of Mathematical Sciences The Johns Hopkins University Baltimore, Maryland 21218	

11. CONTROLLING OFFICE NAME & ADDRESS	12. REPORT DATE
Office of Naval Research Statistics and Probability Program Arlington, Virginia 22217	November, 1985
	13. NUMBER OF PAGES 19

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS
U-statistics, trimmed samples, robust inference, nonparametric, uniform central limit theory.

20. ABSTRACT Let $X_{ni} \leq \ldots \leq X_{nn}$ be an ordered sample of size n. We establish asymptotic normality of U-statistics based on the trimmed sample $X_{n,[\alpha n]+1} \leq \ldots \leq X_{n,n-[\beta n]}$, where $0 < \alpha, \beta < \frac{1}{2}$. This theorem and its multi-sample generalization are illustrated by various statistics of importance for robust estimation of location, dispersion, etc. This unifies the flexibility of the class of U-statistics and the classical principle of rejection of outliers. In addition, as a tool in our treatment, but also having broader interest, a uniform version of the central limit theorem for U-statistics is provided.

END

FILMED

2-86

DTIC